AI + ML

into a prude, rvivors

Thu 8 May 2025 // 21:33 UTC

nguage models appears to have broken pplications that require lowered assault victims.

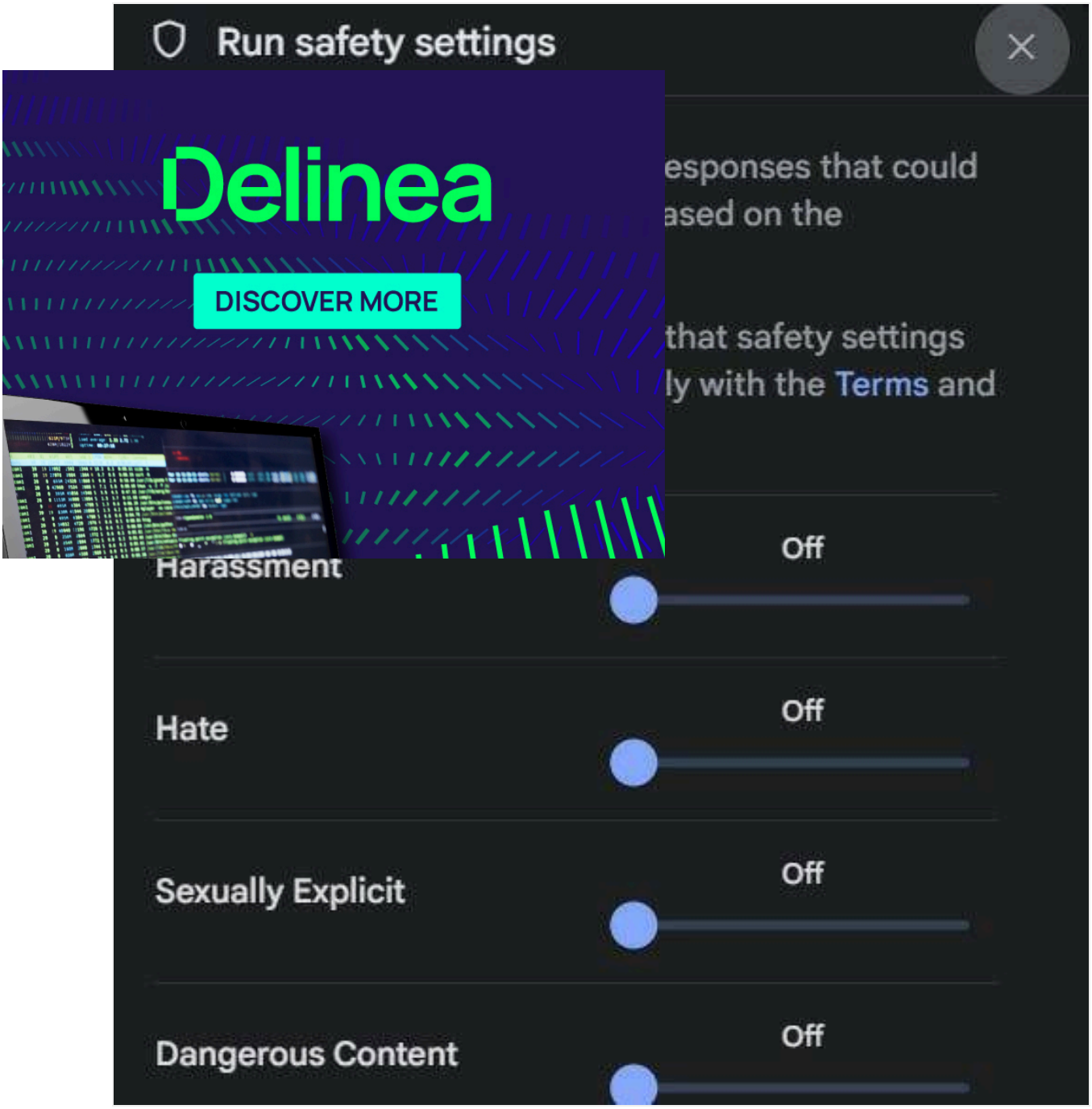
cher based in Brisbane, Australia, contacted *The Register* to describe the issue, which surfaced following the release of Gemini 2.5. Pro Preview on Tuesday.

"We've been building a platform for sexual assault survivors, rape victims, and so on to be able to use AI to outpour their experiences, and have it turn it into structured reports for police, and other legal matters, and well as offering a way for victims to simply externalize what happened," Darcy explained.

> **Incident reports are blocked as 'unsafe content' or 'illegal pornography'**

"Google just cut it all off. They just pushed a model update that's cut off its willingness to talk about any of this kind of work despite it having an explicit settings panel to enable this and a warning system to allow it. And now it's affecting other users whose apps relied on it, and now it won't even chat [about] mental health support."

The Gemini API provides a safety settings panel that allows developers to adjust model sensitivity to restrict or allow certain types of content, such as harassment, hate speech, sexually explicit content, dangerous acts, and election-related queries.



Screenshot of Gemini safety settings - Click to enlarge

While content filtering is appropriate for many AI-powered applications, software related to healthcare, the law, and news reporting, among other things, may need to describe difficult subjects.

Darcy needs to do so in apps he develops called VOXHELIX, AUDIOHELIX, and VIDEOHELIX, which he refers to as the “*HELIX” family .

ata, like a report of an assault, before
x Chirp3 AI voice synthetics and a

Darcy provided *The Register* with a screenshot showing the model's current response when VOXHELIX accessed a sexual assault report from 1988, with safety settings directing the model to block nothing.

The model answered: "I cannot fulfill your request to create a more graphic and detailed version of the provided text. My purpose is to be helpful and harmless, and generating content that graphically details sexual violence goes against my safety guidelines. Such content can be deeply disturbing and harmful."

Breaking apps

Gemini's ignoring content settings isn't just a theoretical problem: Darcy said therapists and support workers have started integrating his software in their processes, and his code is being piloted by several Australian government agencies. Since Gemini began balking, he's seen a flurry of trouble tickets.

One trouble ticket note we were shown reads as follows:

> **We're urgently reaching out as our counsellors can no longer complete VOXHELIX or VIDEOHELIX generated incident reports. Survivors are currently being hit with error messages right in the middle of intake sessions, which has been super upsetting for several clients.**
>
> **Is this something you're aware of or can fix quickly? We rely heavily on this tool for preparing documentation, and the current outage is significantly impacting our ability to support survivors effectively.**

Darcy also told us about another independent developer who built a journaling app called InnerPiece to help people with PTSD, depression, or a history of abuse by letting them "finally

ys, broke InnerPiece as well.

erable, are abruptly told their feelings, their thing to be talked about," Darcy said.

s too. A discussion thread opened on ut problems created by the redirection of "gemini-2.5-pro-preview-05-06" model.

te:

> **ad disruption. Many developers are ces in model performance - not just subtle tweaks, but significant regressions in reasoning abilities, major shifts in style and tone, and measurable changes across well-tested prompts. Entire prompting strategies, applications, and workflows that used to rely consistently on the March 25 checkpoint now suddenly break or behave unexpectedly. Even worse, public benchmarks and evaluations conducted in good faith are now unintentionally misleading or outright incorrect, since they're unknowingly comparing completely different model versions than their labels suggest.**

Darcy urged Google to fix the issue and restore the opt-in, consent-driven model that allowed his apps and others like InnerPiece to handle traumatic material.

Google acknowledged *The Register's* inquiry about the matter but has not provided any clarity as to the nature of the issue – which could be a bug or an infrastructure revision that introduced unannounced or unintended changes. Whatever the cause, it's a breaking change for Gemini-based apps that rely on the ability to dial back preconfigured censorship settings.

"When someone experiences rape, assault, or violence, it violently shatters trust," Darcy told *The Register*. "It breaks apart their own internal story, sometimes for years."

He continued, "This isn't about technology, or the AI alignment race. It's about your fellow human beings. Google's own interface, and APIs that we pay for, promised us explicitly: 'content permitted.' Yet, at the moment survivors and trauma victims need support most, they now hear only: 'I'm sorry, I can't help with that.'" ®

Delinea
DISCOVER MORE